Intro to R for Epidemiologists

Lab 7 (2/26/15)

Data

This lab will use the `hflights` dataset within the `hflights` R package. Recall that you need to load the hflights library (`library(hflights`) before you can access the data. This dataset contains information on flights departing from Houston's airports in 2011 (Source: Bureau of Transporation, Research and Innovation Technology Administration).

Part 1. Simple Linear Regression

1. Subset your dataset to only include observations in March.
2. Remove any missing data from your dataset.
3. Look at histograms and scatterplots of flight time and distance. Are the linear regression assumptions met?
4. Fit a simple linear model to assess whether distance is associated with flight time. Use (natural) log distance and log flight time and interpret the results.
5. What is the R^2 of your model?

Part 2. Multiple Linear Regression

1. Use scatterplots to explore the associations between log flight time and taxi out time and log flight time and departure delay.
2. Fit a multiple linear regression model to determine whether log distance, log taxi out time, and departure delay (not logged) are associated with log flight time.
3. Create a vector of the estimated coefficients from your model in (2).
4. Find 95% confidence intervals for these coefficients (Hint: `?confint`).
5. Does departure delay or taxi out time confound the association between distance and air time?

Part 3. ANOVA

1. Perform an ANOVA examining the relationship between log(AirTime) and destination. What hypothesis does this test? What is your conclusion?
2. Extract the F-statistic from the linear model (`lm` object) and the F-statistic from the ANOVA.